Can Statistics Canada.

(2)

DOMINION BUREAU OF STATISTICS
Census Division
OTTAWA

Reprinted Publication No. 8

COMPUTER METHODS FOR GEOGRAPHICAL CODING AND RETRIEVAL OF DATA IN THE DOMINION BUREAU OF STATISTICS, CANADA

by

I.P. Fellegi and J.I. Weldon

Reprinted from Proceedings of the Social Statistics Section, 1967, Washington, D.C.: American Statistical Association, pp.53-57

Canada, Statistics Canada.

2

DOMINION BUREAU OF STATISTICS
Census Division
OTTAWA

Reprinted Publication No. 8

COMPUTER METHODS FOR GEOGRAPHICAL CODING AND RETRIEVAL OF DATA IN THE DOMINION BUREAU OF STATISTICS, CANADA

by

I.P. Fellegi and J.I. Weldon

Reprinted from Proceedings of the Social Statistics Section, 1967,
Washington, D.C.: American Statistical Association, pp.53-57

COMPUTER METHODS FOR GEOGRAPHICAL CODING AND RETRIEVAL OF DATA IN THE DOMINION BUREAU OF STATISTICS, CANADA

I.P. Fellegi and J.I. Weldon
Dominion Bureau of Statistics

A large up-surge in the collection and use of statistics has been experienced in recent years. It is reasonable to expect that the explosion in statistical activities will continue. We shall have to make sure however that future growth will be controlled, well coordinated and that it will be achieved by efficient utilization of the financial and manpower resources.

In view of these considerations and of the recent technological and scientific developments it is important that developmental work should get underway towards the creation of some general tools applicable to several surveys or data files. Such general tools, such automatic survey systems may represent important means to achieve economies to extend our processing and retrieval capabilities, to enable us to deal with massive volumes of data and to build into our data processing systems important elements of standardization. As such, these general survey systems may be the basic <u>technological</u> prerequisities of large-scale national statistical information systems [13].

The present paper will describe briefly the developmental work underway in the Dominion Bureau of Statistics towards the creation of an automatic geographic coding and retrieval system in larger urban areas. Although the system is expected to be of general utility, we shall discuss it in the context of the 1971 Population Census which, we expect, will be the first large-scale application of it.

The major system features which are to be examined in more detail are as follows:

- data retrieval by user specified areas in larger urban municipalities;
- automatic assignment of geographic location identifiers to urban addresses;
- acceptance and recognition of addresses in free format, automatic correction of the spelling and key punching type errors;
- effective data retrieval and tabulation techniques;
- geocoding and geographic retrieval outside of the large cities;
- reliability of data and disclosure.

Conceptually the function of this system is to retrieve and tabulate geographically coded census data for any arbitrarily defined urban area. Geographic coding is achieved by automatic conversion of addresses to unique geographic coordinates. The entire system operation in a nutshell can be described as follows:

- an address conversion file which can convert urban addresses to unique geographic coordinates is to be produced;
- addresses of the enumerated urban households are to be put into a predetermined standard format, verified and corrected;
- the addresses are to be substituted by their respective geographic coordinates (geocoding);
- the geocoded census data is to be stored for future retrieval;
- tabulations by user specified areas are produced on demand, subject to considerations of statistical reliability and confidentiality.

<u>Data Retrieval by User Specified Areas in Larger Urban Municipalities</u>

In the present context a larger urban municipality is tentatively defined as a city or metropolitan area with a population of 50,000 or over. The important consideration is that a municipality must be a certain size, or part of a large urbanized area to be in a position to take advantage of small area information. The user will be able to delineate on a map the area for which he needs statistical tabulations. Such user-specified areas should preferably not cut through block faces and must be sufficiently large to permit the provision of statistical tabulations without violating the principles of confidentiality. Another problem in connection with statistical small area tabulation which will have to be kept in mind relates to sampling and non-sampling errors.

The user specified retrieval areas are conceived as polygons and are described by the coordinate values of the polygon vertices. Data retrieval for the user specified polygon is done by computer. The programme first selects all the block faces (sides of city blocks between neighbouring street intersections) represented by their midpoint coordinates which are within the user specified area, then retrieves and tabulates the census data for the selected block faces. Characteristically the system approximates the arbitrarily specified areas by using block faces as building blocks. The technique enables us to retrieve by streets or street segments as well [1, 3].

<u>Automatic Assignment of Geographic Location Identifiers to Urban Addresses</u>

This operation is commonly referred to as geocoding. The assignment of geographic coordinates to urban addresses enables us to retrieve by the arbitrarily specified areas. The geographic coordinate of an urban address is that of the block face within which the address is located.

Geocoding is performed with the aid of the address conversion file. This file contains street names, address ranges by block faces and the corresponding block face center point coordinates. The geocoding operation is carried out by

Reprinted from the 1967 Social Statistics Section
Proceedings of the American Statistical Association

providing tabulations by user specified areas is, of course, that the requirements cannot be known in advance, yet the Statistical Bureau has to satisfy these demands without much delay and at a reasonable cost. These restrictions will quite possibly necessitate that census data be organized in random access storage. We also hope to be able to satisfy at least the simpler types of special tabulations by using a generalized, efficient retrieval and tabulation program.

Random Access Storage of the Census Data

The random access file organization appears to hold out several promises for storing and retrieving census data on users' requests which we intend to carefully investigate. If we shall use randomly accessible storage devices we may be able to compress the data to the extent where little storage will be wasted; and we may be able to increase the efficiency of retrieval to the extent that only data required for retrieval would be accessed. This type of file might consist of two modules, which are the data file and the index file. The organization of the records in the data file would be by cities or metropolitan areas and by block faces within them. A record in the census file today typically contains all the characteristics relating to one person. The records of the proposed random access file would be organized by characteristics in a string form, each string containing one of the characteristics for all the enumerated persons. This means that if in a metropolitan area there are one million persons enumerated and there are, say, 50 characteristics reported per person, the proposed file will consist of 50 strings, each one of them one million characters, digits or bits long depending on the data content.

The other file module mentioned was the index file. The index file might be organized in a hierarachical fashion in list mode. The first level of this hierarchy might contain the list of province names and address pointers which are directing to the list of city names within the respective provinces. The second level of the hierarchy might contain the list of city names by provinces and address pointers which are directing to the list of block faces within the respective cities. The third level of the hierarchy contains the list of block faces for each of the cities in the form of block face centroid coordinates and address pointers which are directing to the first sequential appearances of block faces in the various census data characteristic strings.

Retrieval by arbitrary areas can be achieved by listing the coordinate points of the retrieval polygon vertices, accessing the block face centroid list for the requested municipality by descending through the hierarchy of the index file, determining the block face centroids which are contained within the arbitrarily specified retrieval polygon, retrieving the desired characteristic data string portions for the selected block face groups from the census data file, and performing the requested tabulation. The entire operation is an integrated computer process.

Generalized Retrieval Programme

An important aspect in providing fast turn around time at a low cost to users is the availability of a generalized retrieval programme. Input to such a retrieval programme requires the designation of the province, municipality, the listing of the desired characteristics and retrieval conditions for tabulation or cross-tabulation, and the coordinate points of the vertices for the requested retrieval polygon. The significance of such a generalized programme would be that at least simpler types of special tabulations could be specified through the use of the programme without extensive training in programming. The data file organization by characteristic strings and the index file organization in hierarchical structure would greatly facilitate the utilization of such a generalized programme. The most significant advantage of such high level retrieval languages is that they permit the description of the retrieval and tabulation requests in some restricted English language form, which then can be used as an input to a computer programme to perform the designated operations.

The system must also be designed to facilitate an inverse retrieval function. This refers to the type of request which seeks the delineation of an area (or areas) which satisfy some stated conditions. After having determined the desired area its boundary could be mapped by means of computer graphics.

The Problem of Geocoding and Geographic Retrieval Outside the Larger Cities

Automatic geocoding assigns to postal addresses, with a minimum of manual intervention, the location-specific coordinates of the center point of the block face in which the address is located. In this fashion the traditional coding is carried out in that the address is identified as belonging to a particular pre-designated standard area, the block face in the present case. Automatic geocoding differs, however, from the traditional geographic coding in three important ways. First, it carries the coding to much smaller areas (block faces) than would be conceivable using manual methods, hence it provides very small building blocks for future aggregations. Second, it provides a reasonably error-free general tool that can be applied to data files, whatever their origin, as long as the data fields are identified

[5] Fellegi, I. P., "Response and its estimation", *Journal of the American Statistical Association*, 59 (1964), 1016-1041.

[6] Fellegi, I. P., and Krotki, K. J., "The Testing Programme for the 1971 Census in Canada", *Proceedings of the Social Statistics Section*, *American Statistical Association* (1967).

[7] Hansen, M. H., Hurwitz, W. N., and Bershad, M. A., "Measurement errors in censuses and surveys", *Bulletin of the International Statistical Institute*, 32nd Session 38 (1959), 359-74.

[8] Hansen, M. H., Hurwitz, W. N., and Pritzker, L., "The estimation and interpretation of gross differences and the simple response variance", *Unpublished report*, *Bureau of the Census*, U.S.A. (1963), (honouring Professor P. C. Mahalanobis).

[9] Hanson, R. H., and Marks, E. S., "Influence of the interviewer on the accuracy of survey results", *Journal of the American Statistical Association*, 53 (1958), 635-55.

[10] Kish, L., and Lansing, J. B., "Response errors in estimating the value of homes", *Journal of the American Statistical Association*, 49 (1954), 520-38.

[11] Kish, L., "Studies of interviewer variance for attitudinal variables", *Journal of the American Statistical Association*, 57 (1962), 92-115.

[12] Mahalanobis, P. C., "Recent experiments in statistical sampling in the Indian Statistical Institute", *Journal of the Royal Statistical Society*, 109 (1946), 325-70.

[13] Nordbotten, S., "Automatic files in statistical systems".

[14] Pritzker, L., and Hanson, R. H., "Measurement errors in the 1960 Census of Population", *Proceedings of the Social Statistics Section*, *American Statistical Association* (1962), 80-90.

[15] Sukhatme, P. V., *Sampling Theory of Surveys with Applications*, Ames, Iowa: Iowa State University Press and New Delhi, India: *Indian Society of Agricultural Statistics*, 1954. Pp. 445-6.

[16] Sukhatme, P. V., and Seth, G. R., "Non-sampling errors in surveys", *Journal of the Indian Society of Agricultural Statistics*, 4 (1952), 5-41.